Master in

Actuarial Science

Master's Final Work

Internship Report

Health Insurance Pricing with Generalised Linear Models

Ana Beatriz Marques Cabral Valente

October - 2020

LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

Master in

Actuarial Science

Master's Final Work

Internship Report

Health Insurance Pricing with Generalised Linear Models

Ana Beatriz Marques Cabral Valente

Supervision:

João Manuel de Sousa Andrade e Silva

Maria Isabel Teixeira da Silva Pimenta Ribeiro

October - 2020

Acknowledgments

The realization of this internship report was a pleasant challenge and this accomplishment was only possible with all the support I received.

First of all, I would like to express my deepest appreciation to my thesis advisor, professor João Andrade e Silva, who provided valuable and constructive suggestions. His extensive knowledge and commitment were essential for the materialization of this report.

I would also like to acknowledge my colleagues at Allianz for their patient guidance and assistance. Despite all the barriers created by the pandemic situation and the home confinement, your support was fundamental and is greatly cherished. Special gratitude should be given to my company's supervisor, Isabel Ribeiro, for embracing this project by my side with an enthusiastic encouragement present at all times.

Finally, I must express my gratitude for the unconditional love of my family, friends and dogs. Thank you for always being there, your moral support and confort throughout this confinement phase were necessary to maintain the focus and finish this report.

Abstract

Generalized Linear Models (GLMs) are being broadly used in the Non-Life Insurance Pricing. The premium charged by the insurance company is calculated based on a tariff. The most standard procedure to estimate the pure premium is by assuming that the claim counts and claim amounts are independent. From this independence, the claim frequency and severity can be forecasted by distinct GLMs and the Tariff is obtained by combining both models.

The present report gives a brief introduction on the methodology and describes how we prepared the data prior to the GLM application. The models obtained for the Stomatology Treatments and Appointments, one of the many coverages that can be included in a Health Insurance policy, are analyzed in this report. The SAS software was used to construct the datasets and to properly organize the data and R was the software used for the modelling process. Once the models were estimated, the pure premium was calculated and a tariff for the mentioned coverage was constructed.

Finally, we compared the results obtained by modelling the coverage in R with the output obtained by my colleagues, using the software implemented by the company. We conclude that both models are not significantly different, despite having some structural distinctions.

Keywords: Health Insurance, Insurance Pricing, Tariff, Generalized Linear Models, Claim Frequency, Claim Severity

Resumo

Os Modelos Lineares Generalizados (GLMs) são amplamente utilizados na precificação de seguros do ramo Não Vida. O prémio cobrado pela seguradora é calculado com base em uma tarifa. A abordagem clássica para estimar o prémio é feita assumindo a independência entre o número de sinistros e o seu custo. A partir desta independência, a frequência e a severidade dos sinistros são estimados através de GLMs separados e a tarifa é obtida combinando os dois modelos.

O presente relatório fornece uma breve introdução sobre a metodologia e descreve como preparámos os dados antes da aplicação do GLM. Os modelos obtidos para os Tratamentos e Consultas de Estomatologia, uma das muitas coberturas que podem ser incluídas numa apólice de Seguro Saúde, são analisados neste relatório. O software SAS foi utilizado para construir as bases de dados e para organizar adequadamente a informação e o software R foi utilizado para o processo de modelagem. Uma vez estimados os modelos, o prémio puro foi calculado e a tarifa, para a cobertura mencionada, foi construída.

Por fim, comparámos os resultados obtidos em R com as conclusões obtidas pelos meus colegas, utilizando o software implementado pela empresa. Concluímos que ambos os modelos não são significativamente diferentes, apesar de apresentarem algumas distinções estruturais.

Palavras-Chave: Seguro de Saúde, Precificação de Seguros, Tarifa, Modelo Linear Generalizado, Frequência de Sinistros, Severidade de Sinistros

List of Tables

List of Figures

Contents

Chapter 1

Introduction

1.1 Motivation and Goals

The present internship report is the result of work developed from February to August at the insurance company Allianz Portugal. I integrated the Non-Life Actuarial Department, that is accountable for the elaboration and management of the pricing of each product. It is also responsible for the production and analysis of forward-looking Key Performance Indicators (Kpis) based on the technical premium and for the identification of profitable growth opportunities.

The internship was centered on the elaboration of a tariff for group and individual health insurance with the possibility of creating a simulator as time permits. The main activities were:

- Data building and analysis, using SAS programming.
- Construction of a profile, to have an early idea of how the frequency, the average cost and the risk premium are distributed through the years and through different features like age, districts, methods of payment among others.
- Frequency and Severity GLM modelling for each coverage, using Emblem software.
- Creation of a Pricing Simulator in Excel.
- Improving the tariff in use.

As a result of the pandemic crisis we are going through, my internship had to be done under home confinement. This unexpected event launched some difficulties in the

performance of our project, hence some adjustments had to be made to cope with the situation.

The biggest obstacle was the impossibility of learning how to use the Emblem software, which is implemented by the company to perform the GLM modelling. We ended up choosing one health coverage to model in R and then comparing the results with my colleagues that used the assigned program.

Not having the full internship experience was an unfortunate event and the productivity of working from home was not the same as being in the company. Therefore, we had to extend the internship and, even then, we were unable to finish the project as agreed. Despite all the setbacks, the internship represents a very important learning moment and it was a pleasure to be part of this project.

1.2 Context

A Health Insurance policy is a contract celebrated between the insurer and the customer, where the former is committed to assume various health related expenses, suffered by the latter, for a given price and during a certain time period. There are two types of health insurance: Group and Individual.

The Group insurance provides healthcare coverage to a party of individuals that belong to the same organization. Depending on the chosen product, the group members have the option to accept it and often even extend the coverage to their family members. This type of insurance is purchased and usually, partially covered by the organization.

The Individual health insurance is obtained directly by the individual and allows him to select a health insurance coverage plan with a wider range of features that can be customized to his demand and on a family basis. Therefore, this type of insurance is usually more expensive.

Different health plans include coverage for hospitalization and may also cover for the ambulatory services (appointments, treatments, exams), childbirth, stomatology, prescription drugs, prosthesis and orthosis, dental and others. Depending on the health plan selected by the policyholder, the personal characteristics and on the features of the contract, the insurance company needs to charge a value per individual insured, in return for protection against financial losses during a time period, customarily of one year. This fee

is called premium.

The premium is calculated considering the expected loss that the insurance company is accepting to cover. However, it is impossible to know exactly how many claims a customer will have in the following year (claim frequency) or how expensive are they going to be (claim severity). For many years, actuaries have been facing this uncertainty. Many statistical modelling tools and techniques were developed to help estimate accurately all the expected costs and improving the accuracy of the predictions. Therefore, with the increased volume of data, insurers can appraise risks more extensively outreaching better predictions and create prices more adequately.

To this end, a well-developed class of predictive models, called Generalized Linear Models (GLM), is implemented by insurance companies to estimate accurately the expected losses.

The present internship report will explore the application of GLM on modelling the premium by means of estimating separately the claims frequency and severity, as well as all the challenges encountered while dealing with huge amounts of data. For confidentiality reasons, the data presented in the report was obtained by using a subset of the company's data, not representing the entire portfolio.

1.3 Report Organization

This internship report has the following structure. In chapter 2 a brief theoretical background on insurance ratemaking and GLM is presented. In chapter 3 is succinctly illustrated how we prepared the data used in this project and the relevant variables are introduced. In chapter 4 the models for the Claim Frequency and Severity are displayed along with the final model for the Pure Premium. Finally, chapter 5 comprises the comparison between the obtained results and possible further developments.

Chapter 2

Basic Concepts of Non-Life Insurance Pricing

In this chapter, we describe how to calculate the insurance premium based on a tariff built from historical data. All the analyses done for this purpose were based on Generalized Linear Models (GLM). Hence, we will also give a brief introduction to this class of models.

2.1 Premium Concepts

It is commonly known in ratemaking that the costs related to an insurance contract i, in the following year, can be jointly composed by two variables that are generally assumed independent:

1. The number of claims incurred in policy i per exposure unit, N_i.
2. The claim severity X_{ij} that represents the total loss amount of the j^{th} incurred claim by policy i, where $j = 1, ..., N_i$.

The individual risk model compose the aggregate loss (S_i) as the sum of the total loss amount from the policy i, hence $S_i = \sum_{j=0}^{N_i} X_{ij}$, where $X_{i0} \equiv 0$. Conditionally on N_i, the random variables $X_{i1}, ..., X_{iN_i}$ are assumed to be independent.

The individual expected loss, also known as the individual pure premium, is then equal to

$$E[S_i] = E[N_i] \cdot E[X_i] = \mu_{N_i} \times \mu_{X_i}, \tag{2.1}$$

where $E[N_i]$ is the expected number of claims per unit of exposure and $E[X_i]$ represents the expected claim severity for policy i.

Moreover, the variance of the individual expected loss can be computed by taking advantage of the iterated expectations:

$$\begin{aligned} Var(S_i) &= Var\left(E\left[N_i\bar{X}_i|N_i\right]\right) + E\left[Var(N_i\bar{X}_i|N_i)\right] \\ &= (\mu_{X_i})^2\, Var(N_i) + E\left[N_i^2 Var(\bar{X}_i|N_i)\right] \\ &= (\mu_{X_i})^2\, Var(N_i) + \mu_{N_i} Var(X_i), \end{aligned} \tag{2.2}$$

where $\bar{X}_i$ is the average cost of claims given that a claim occurred, $\bar{X}_i = \frac{\sum_{j=1}^{N_i} X_{ij}}{N_i}$.

The total loss amount for a given portfolio with n policies corresponds to the sum of all the policy losses, $S = \sum_{i=1}^{n} S_i$. By considering independence across the loss amounts $S_1, ..., Sn$, the variance of the total loss can be defined as the sum of all the variances. Thereupon, the expected total loss amount and its variance are represented respectively by

$$\mu_S = \sum_{i=1}^{n} E[S_i] \tag{2.3}$$

$$\sigma_S^2 = \sum_{i=1}^{n} Var[S_i] \tag{2.4}$$

However, the pure premium represents only part of the overall price paid for the insurance product. The total premium charged to the policyholder, known as the gross premium, comprises the pure premium, a value to cover for the risk taken by the insurance company together with any other expenses and loadings for profit.

2.2 Tariff

In Insurance, the premium charged is calculated based on a tariff. The development of a tariff is an actuarial study based on the claims historical records and policy level information.

The most standard procedure to estimate the pure premium is by assuming that the claim counts and claim amounts are independent random variables. From this independence, the claim frequency and severity can be forecasted by distinct GLMs, and the expected loss is obtained by the multiplication of the two mean estimates. This estimation is based on how the dependent variable varies with a set of predictors, called rating

factors or covariates. An advantage of this tariff model is the possibility of assessing the different effects that the rating variables have in each model. If we were directly modelling the pure premium, we would not be able to see many of these effects.

The rating factors considered for calculating the premium for a certain insurance product vary, and usually they can be:

- Features of the policy contract: age of the policy, method of payment (monthly, quarterly, semiannual, annual), type of insurance (group or individual), etc.
- Characteristics of the insured person: age, sex, geographic region, income, etc.

Generalized Linear Models are widely used in the insurance business for forecasting the pure premium and are going to be introduced further ahead. This type of models usually categorizes the quantitative and qualitative rating variables into several intervals and buckets, respectively, so the tariff has a simple structure and can be easily implemented.

If two or more policyholders have similar risk profiles, i.e. if they are in the same interval or bucket for every rating variable , we say that they belong to the same tariff class and they are going to be charged with the same premium.

2.3 GLM Overview

Generalised Linear Models (GLM) are a very acquainted topic nowadays, for this reason we will present briefly a theoretical introduction following Ohlsson E. (2010) and Goldburd et al. (2016).

To estimate the individual's pure premium, the insurer deploys predictors, also known as covariates or explanatory variables. These predictors, as mentioned before, are typically policyholder's features or any other insured product's characteristic that the insurer thinks is beneficial to be used in the tariff. The GLM allows to capture the effects these predictors have on the target variable, the one we are interested in.

Usually, GLM has three fundamental components:

1. It is assumed that the dependent random variable Y follows a distribution from the exponential dispersion family. Under the exponential family fall all the distributions that can be generalised from the following canonical form, for the i^{th} observation:

$$f(y_i;\theta_i;\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}, \tag{2.5}$$

where , $a(.)$, $b(.)$ and $c(.)$ are known functions, θ_i is the canonical parameter, ϕ is the constant dispersion parameter and the mean-value parameter is denoted by μ_i.

Distributions like Poisson, Gamma, Normal and many others belong to the exponential dispersion family and can be written in the canonical format (2.5), (see Dobson & Barnett, 2002, among others).

For this family of distributions, the mean and the variance of the response variable can be obtained by

$$E[Y_i] = b'(\theta_i) = \mu_i \tag{2.6}$$

$$Var(Y_i) = b''(\theta_i)a(\phi) = V(\mu_i)a(\phi) \tag{2.7}$$

Note in equation (2.6), that θ_i is an invertible function of μ_i , allowing us to define $\theta_i = (b')^{-1}(\mu_i)$. On equation (2.7) , $V(\mu_i)$ is called variance function and it describes how the variance depends on the mean.

2. A linear predictor η_i, that is represented by an intercept β_0 and a linear combination of the p covariates x_{ij} and the regression parameters $\beta_1, ..., \beta_p$.

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{2.8}$$

where $\beta_0, ..., \beta_p$ are unknown parameters that must be estimated by the historical data.

3. A link function $g(\mu_i)$, that engages the linear prediction η_i with the expected value of the target variable, μ_i :

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{2.9}$$

This function is usually selected by the modeler and it needs to be strictly monotonic and differentiable with an inverse function $g^{-1}(\mu_i)$. Although the link function is useful for the model specification, the value we want to predict is μ_i. To derive μ_i from the link function one just has to apply the inverse function, $\mu_i = g^{-1}(\eta_i)$.

Frequently, in tariff estimation, the link function used to estimate the claim frequency and severity is the log-link function. By applying this function, we produce a multiplicative model represented by equation (2.10), from where the value of μ_i can be obtained by solving equation (2.11).

$$\ln(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{2.10}$$

$$\mu_i = e^{\beta_0} \times e^{\beta_1 x_{1i}} \times \cdots \times e^{\beta_i x_{ip}} \tag{2.11}$$

The multiplicative model is easy to implement and very intuitive. A variation in one of the covariates, *ceteris paribus*, influences proportionally the value of the linear predictor. The log link function also works well with the distributions we are going to use later on, Poisson and Gamma distributions, and ensures that we cannot attain negative fitted values.

After obtaining the estimates for all the coefficients of the frequency and severity models, we are able to calculate the μ_{N_i} and μ_{X_i}, and by applying equation (2.1) we obtain μ_{S_i}. Another advantage from this type of rating structure, is that by multiplying two multiplicative models, we obtain one with the same structure.

2.4 Exposure, Offset and Weight

Exposure is a measure of weight attached to the dataset that allows all observation values to be comparable, after adjusted by that weight. The exposure depends heavily on the variable we aim to model and can be inserted in the GLM as an offset or a weight variable.

In insurance rating, when dealing with data exposed to the risk for a different length of time or that varies with any other exposure unit, it is necessary to make an adjustment. By modelling with GLM it is possible to overcome this difference by adding to equation (2.9) a predictor variable called offset. The offset corresponds to an adjustment to the mean:

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \text{offset}_i \tag{2.12}$$

Sometimes, the data we are dealing with is aggregated and we have only access to the average amount instead of all the individual information. For example, if we are modelling

the claim severity, each row in our dataset will be referring to the average loss amount of all the claims belonging to that row.

To deal with this kind of situation, we can take advantage of a weight variable ω_i, that gives a weight to each observation during the estimation process. The records that have more exposure to the risk, i.e. rely on more data, should have a bigger weight on the estimation. In the previous example, the weight variable is the number of claims aggregated per row, this in turn means, that a row representing the average loss amount of ten claims should have twice the weight during the estimation procedure than a row aggregating only five claims.

Mathematically, the weight variable is introduced in the model as an adjustment to the variance. So, the variance for observation i, specified in equation (2.7), can be rewritten in the presence of ω_i as the following:

$$Var(y_i) = \frac{V(\mu)a(\phi)}{\omega_i} \tag{2.13}$$

Sometimes, is not straightforward to know whether we should apply the offset or add the weight variable. Moreover, there are cases where, depending on the manner in which these are included, we can get the same final results when a weight or an offset are considered. In the next example we present a scenario where we can get equal estimation values by applying both adjustments.

When modelling the number of claims per policy it is important to deem as exposure the length of time each policy was in force. For instance, if in a Homogeneous Poisson Process we have two policies, one with an entire year of exposure and the other with four months of exposure, the first policy will have thrice the expected number of claims, *ceteris paribus*. Therefore, in the case where we have a claim count model, the greater the exposure the greater the expected number of claims. Hence, it would be necessary to take into consideration the exposure time. When using a log link function, the offset will be equal to the log of exposure. By adding this offset, we are increasing the mean of the number of claims proportionally to the exposure, that will result in an increase of the variance when using a Poisson distribution. So, in the present situation, it is not necessary to adjust the variance by adding a weight variable.

Alternatively, if we were modelling the ratio between the number of claims and the exposure (claim frequency), the mean is expected to remain constant, but the variance

would decrease because now we are considering a variable with higher stability. Thereupon, we need to add the weight variable to adjust the variance.

The equivalence between implementing the exposure in the offset, when modelling the number of claims, or adding it as a weight variable, when modelling the claim frequency, is succinct in Table 3.1, taken from Goldburd et al. (2016) :

	Claim Count	**Frequency**
Target Variable	Number of claims	$\frac{\text{Number of claims}}{\text{Exposure}}$
Distribution	Poisson	Poisson
Link	log	log
Weight	-	Exposure
Offset	log (Exposure)	-

Table 2.1: Equivalence between implementing the exposure in the offset or adding it as a weight variable, when modelling the number of claims and the claim frequency, respectively.

2.5 Goodness of fit

The aim of model fitting is to find a parsimonious model that fits the data well, i.e. a model where the fitted values $\hat{\mu}$ are as close as possible to the observed data y. In the model building process, presented in the next chapters, we will be using several well-known statistical measures that are useful for analysing the importance of each variable and compare different models.

2.5.1 Log-Likelihood and Deviance

The GLM is developed by using the Maximum Likelihood Estimation (MLE) method. As mentioned before, all the exponential family distributions have a density function $f_Y(y_i; \theta_i; \phi)$ with the canonical form (2.5). By assuming independence between the observations, if we multiply the density functions that have been assigned to each observation i, we obtain the Likelihood function. As usual, we will work with the log of the likelihood.

The log-likelihood function, for the exponential dispersion family, is given by:

$$l(\boldsymbol{\theta};\boldsymbol{Y}) = \sum_{i=1}^{n} \left[\frac{Y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c_i(Y_i,\phi) \right], \tag{2.14}$$

where $\boldsymbol{Y}$ represents the vector of all response observations and $\boldsymbol{\theta}$ is the vector of the n canonical parameters, $\theta_1, ...\theta_n$. The GLM is then fitted by encountering the set of coefficients β that maximize the log-likelihood function $l(\boldsymbol{\theta};\boldsymbol{Y})$.

One way of appraising the performance of a model is by comparing it with the saturated model. The saturated model is a more general model with the same number of covariates as data observation points. Despite of being an overparameterized model, it gives us the highest value possible for the log-likelihood function. Therefore, the deviance compares the log-likelihood of our model of interest, ll_{model}, with the log-likelihood of the saturated model, $ll_{\text{saturated}}$:

$$\mathrm{D}^\star = 2 \cdot (ll_{\text{saturated}} - ll_{\text{model}}), \tag{2.15}$$

where $D^\star$ represent the scaled deviance.

The deviance measures how close the fitted values are to the observed data, the smaller the $D^\star$, the better the model fits the data.

However, the deviance has a limitation. This measure can only be applied for comparing models if they are nested. Two models are nested if one is the subset of the other, i.e. one model can be obtained from the other by a adding a set of linear constraints.

Another important result is the null deviance. It gives us the scaled deviance for the model with no predictors, acquainted as null model. The null deviance works as an indicator of how much the existence of the simplest model adds to its absence. Every model can be derived from the null model, therefore we can always compare the deviance of the resulting model with the null deviance to analyse how much adding a certain linear restriction improved the model. In R, after building a GLM, we obtain in the output the null and the residual deviance.

2.5.2 AIC and BIC

A widely used measure, to compare non-nested models, is the Akaike Information Criterion (AIC). This model prevents the overparameterization by using the number of pa-

rameters to penalize the complexity of the model.

$$\text{AIC} = -2 \cdot [ll_{\text{model}} - p] \tag{2.16}$$

As for the deviance, the smaller the AIC the better our model fits the data.

Another measure that incorporates the log-likelihood is the Bayesian information Criterion (BIC).

$$\text{BIC} = -2 \cdot ll_{\text{model}} + p \log(n) \tag{2.17}$$

Similarly to AIC, BIC has a penalty parameter but it also includes the logarithm of the number of observations n. A smaller BIC indicates a better model fit.

This two measures not only help find a model that fits the data well, but also a model that is parsimonious.

2.5.3 Test MSE

The Mean Squared Error (MSE) is another statistical measure used to check the accuracy of the model. The MSE is calculated by the following formula:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{2.18}$$

Where n represents the total number of observations, y_i is the actual value for the i^{th} observation and $\hat{y}_i$ is the fitted value for the i^{th} observation. Basically, we are just calculating the mean of all the squared differences between the actual and the predicted values for each observation i. The smaller the MSE, the better are the overall predictions.

The MSE is also implemented by the Cross-Validation approach that will be discussed in the next chapter.

Chapter 3

Data Processing

In this chapter we start by introducing the data and how we obtained the final dataset. We also describe how we split this dataset in two parts by using two different approaches: the Train and Test split and the Cross Validation. Afterwards, we introduce the coverage we are going to model and the variables that are important for our models.

3.1 Main datasets

The data used for the entire project was provided by the company and it contains policyholder-level information from 2016 until 2019. Two main datasets were built for the study. The first dataset contains the exposure-level premium and all the relevant information for each insured person. The second one, comprises all the data on the claims incurred during the period of interest.

Both datasets were merged, obtaining a final dataset, properly organized and without any erroneous information, where we matched each claim with the correspondent policy record.

Prior to any aggregation, we had to work in each raw dataset individually in order to prepare the data accordingly to our goals. This step was time consuming and a big challenge. We had to clean the datasets from double records and misinformation, particularly we eliminated observations with risk exposure equal to 0. During the entire process, we made constant validations to check that we were not losing any important information. We created variables from the available information, such as the age of the policy and the exposure time, and others suffered substantial changes. Most variables were converted into factors because the software used by the company, Emblem, requires that format.

However, by doing so, we gained higher flexibility and obtained a nonlinear structure. For the continuous variables, we grouped them by intervals, already established by the company, and treated them as categorical.

When merging the two datasets, we faced the problem of having several claims, related to the same person, on the same day e.g. appointments or exams. According with the procedure in force by the company, those claims were aggregated and we added their costs. By doing so, we are interfering with the variability and we are losing granularity - for instance, if a particular insured person makes two different exams on the same day, costing 5 € and 35 € each, by aggregating them we would only know that 2 claims occurred and their total cost was 40 €.

3.2 Partitioning the data

When modelling health insurance, it is wise to model each coverage separately. So, instead of using the final dataset that takes into consideration all the existing policies, we created smaller datasets with the policies related to each coverage. When the coverage we are modelling has sub coverages, it is also highly recommended to subdivide the data related to each sub-coverage and model it separately. For example, the ambulatory services would be divided in appointments, urgent care, treatments and exams.

Before commencing on the model building process, we need to perform another partition of the data that will be considered until we find the suitable model. We will use two different approaches to splitting the data: the Train and Test split and the Cross Validation.

3.2.1 Train and Test

The train and test technique randomly splits the dataset into two parts. One part is known as the training set, it represents 80% of the data and is used for model building purposes. The other part is commonly called as test set and, as the name implies, it is required to test the model potential after it has been trained and to compare the candidate models. The ratios used for the train and test were chosen by the company, and they are the typical ones implemented when a big amount of data is available.

Firstly, we make a selection of the models that perform well in the training set and

only then the test set is used to compare the candidate models and choose the one that best fits the data.

The reason for the need of a holdout set of data – the test set – is to avoid overfitting. If the model is selected by only training and analysing its performance in the same dataset, we might get parameters that are good enough only to estimate that set of data, which means that model would not be a good fit for a different set of data. To overcome this bias situation, we must test our model on data that the model has never seen before.

3.2.2 Cross Validation

A different technique that can be implemented to split and evaluate statistical models is the Cross Validation method (CV). Cross Validation has several approaches and all of them involve an iterative process of evaluating the model, making maximal use of the available data. The one we are going to discuss is the k-fold cross validation.

In the k-fold Cross Validation, the data is randomly divided into k groups, also known as folds. The general procedure, for each of the k-folds, is described as:

- Save that fold as the test set.
- Use the remaining $k-1$ folds as the training set.
- Train the model in the training set and asses its performance in the test set.

In this iterative process, every fold serves as the holdout set once. Hence, contrarily to the Train and Test approach, the Cross Validation will end up by using the entire dataset in the iterative process. For each iteration we obtain the prediction error. At the end, we have a value called cross validation error that corresponds to the average of those k prediction errors, the MSE.

The k value is chosen by the user. It is good to keep in mind that the variance of the resulting prediction decreases as the value of k increases. However, the bigger k the more iterations have to be computed, hence more time is needed.

The Train and Test partition technique is the approach implemented by the company where I did my internship. Since I ended up doing a model by myself in R, I splitted the data in a training and test set and I took the freedom of implementing the k-fold Cross Validation in the training set. The Cross Validation was only used to help on the selection

between candidate models, always returning to the original training set to perform the model building process.

3.3 Data Overview

The coverage selected to be modelled in R was Stomatology. This particular coverage, as many others, is divided into sub-coverages: "Treatments and Appointments" and "Prosthesis". As mentioned before, each one of this sub-coverage needs to be modelled separately. The one whose the results are going to be presented further is the Stomatology Treatments and Appointments. It is important to mention that Stomatology is a type of coverage only available for group insurance, hence all the appurtenant data we are going through in this section refers to individuals who belong to group policies.

For confidentiality reasons, the following data was obtained by using a subset of the company's data, not representing the entire portfolio. The dataset used for this sub-coverage englobes, for the calendar years 2016-2019, a total of 26 656 different insured people and 21 759.9 years of exposure. For each observation, we know the number of claims, the total claim cost, the exposure time per year and we have a set of 53 rating factors.

Year	2016	2017	2018	2019	Total
Number of people insured	4008	6691	7686	8271	26656
Exposure	3039	5541.8	6200	6979.1	21759.9
Number of claims	1426	2863	3296	3685	11270
Total Cost (€)	53231.3	103264.6	115636.2	130774.6	402906.7
Frequency	0.47	0.52	0.53	0.53	0.52
Average Cost	37.33	36.07	35.08	35.49	35.75

Table 3.1: Summary of the number of observations per calendar year.

In Table 3.1 we aggregated the total number of people insured, the years of exposure, the number of claims observed and the total cost per year. An increase of the number of insured throughout the years can be observed which, consequently, leads to an increase in the observed number of claims incurred and in the total cost. We also calculated the

claim frequency and the average cost per claim. From Table 3.1, we can observe a growth in the frequency after the first year, where it remains stable, and an average cost per claim also relatively stable over the years, so it was not necessary to deal with inflation.

Number of claims	0	1	2	3	4	≥ 5
Observations	20543	3427	1456	622	294	314
	(77%)	(13%)	(5.5%)	(2.3%)	(1.1%)	(1.1%)
Exposure	15980	3196.6	1394.9	595.9	287.4	305.1
	(73.4%)	(14.7%)	(6.4%)	(2.8%)	(1.3%)	(1.4%)
Total Cost (€)	0	156202	104102.7	59749	37770.34	45083.41
Average Cost per claim	0	45.58	35.75	32.02	32.12	23.87

Table 3.2: Absolute and relative frequency of the number of people, total cost and average claim cost, per claim number.

For the entire observation period, we aggregated the number of people insured, the years of exposure, the total cost and the average cost per claim by the number of claims incurred. The results we obtained are displayed in Table 3.2. We can conclude that, during the period their policies were in force, 77% of the insured did not report any claim and they correspond to 73.4% of the total years of exposure. Some of the insured filled one claim, and 10% reported two or more claims. That gives a total of 11 270 claims observed and a total loss of 402 907€. It is also displayed in the table, that the average cost per claim decreases as more claims are reported per calendar year.

3.4 Variables

In this section we introduce the techniques used for the variable's selection and treatment. We are also going to present the relevant variables that are going to be implemented in our models.

3.4.1 Variables Selection

The response variables, also known as target variables, are the ones whose expected values we are interested in estimating. After choosing the target variable and selecting the suitable distribution, an important step in model building is the selection of the covariates.

To begin, we discarded the variables that, at first sight, are not relevant for modelling Stomatology.

Then, the optimal model variable structure is obtained by doing a thorough classification of all possible variables. We started by assessing the performance of each variable or interaction of variables individually. For that, we constructed single factor models (models with only one variable or interaction) and we ascertained the predictive power of that only variable. We decided to keep all significant variables for a threshold of 5% significance level. At this point we were able to reduce our initial 53 rating variables, but we are still seeking for our desired parsimonious model.

We also performed some descriptive statistics and one-way tables that where useful to have a practical insight within each factor variable, such as the mean, mode, frequency, average costs and risk premium.

Afterwards, with all the significant predictors, we built a "saturated" model and applied a function, existent in R, called ***stepAIC*** from the MASS package. This function goes step-by-step analysing how the AIC varies with the introduction/extraction of each variable in the model. If the variable increases the AIC, then is removed, otherwise is added back to the model. We also made use of other R tools like the cross validation, MSE test, anova test and graphical illustrations, to help us selecting the key variables.

The variables that look significant for the single factor model might not bring any marginal value when inserted in a model with more variables, and the opposite can also occur. We need to take into account the possibility of collinearity between the predictors. Thereupon, with the final significant explanatory factors, we calculated the variance inflation factors (VIF) for a linear model comprising all those potential predictors. VIF measures how much the correlation among the predictors inflates the precision of the estimation. We obtained the VIF by means of the R function called ***VIF***, from the R package car. The rule of thumb used with this measure was that for values above 4, the variable needs to be investigated further and for values above 10, it suffers from extreme collinearity.

3.4.2 Grouping Categorical Variables

Each categorical variable can take on a certain number of possible values, known as levels. Some variables have too many levels that result in too many unnecessary degrees

of freedom, which need to be grouped and revised before including them in the model.

For instance, we transformed the continuous variable *age* into categorical, and obtained a factor variable with too many levels. Since we are referring to a variable with an inherent order, the best way to decrease the number of levels is by grouping neighboured levels that have similar behaviour towards the response variable. Hence, we combined levels based on their coefficients or empirical frequencies.

When we want to group a variable that does not have a specific order, it is more challenging to know which levels to combine. Usually, one starts by grouping the levels with small volume of data or that are not statistically different from the base level. For that, we combine the variables that have similar coefficients. However, that process may be very time consuming for variables with a large number of levels.

Gladly, there are some tools in R that can help us on deciding how to group the levels. One helpful function is the ***fit.contrast***, from package gmodels. With this function it is possible to test for factor variables, with 3 or more levels, if two levels have a similar effect towards the base level and can be grouped, by comparing the resemblance between the coefficients obtained from the GLM model. Another helpful tool is to look at the standard errors of the coefficients (SE), that give us a measure of precision of the estimation. The larger the SE, the less accurate the estimation is.

3.4.3 Variables Introduction

In the context of this internship report, the response variables we are going to analyze are the number of claims observed during the insured period, *ncl*, and the total loss incurred per claim, *severity*.

The explanatory variables that passed the majority of the tests previously described, are introduced in the table 3.2. These are also the variables that are going to be implemented in the models presented in the next chapter.

Variable	Description
Age	The age of the person insured.
Bin_pills	Binary variable that indicates if the policy also has coverage for prescription drugs or not.
Cp2_company	The first two postal code digits of the place where the policyholder works.
Dist_company	The district ("distrito") of the place where the policyholder works.
Exp	Value between (0,1] that represents the amount of time the policy was in force during each calendar year.
Fcob	Indicates if the payment is made by bank transfer or through an agent.
Fleet	Indicates if the company covers in full the premium of their workers (E) or not (M).
Fp	Represents the payment method: monthly, quarterly, semiannual, annual.
Kinship	The relationship the insured person has towards the policyholder.
Policy_age	The age of the policy.
Sex[1]	Gender of the person insured.

Table 3.3: Description of the variables used in our GLMs.

[1]In December 2012, the European Court of justice implemented a rule to defend gender equality in insurance pricing, where is not allowed to price an insurance product based on the gender. In this report the sex variable is only used for internal analysis.

Chapter 4

Models

This chapter comprises the final GLM models obtained from the implementation of the techniques described previously.

4.1 Claim Frequency

For modelling Claim frequency, the target variable we considered was the number of claims, *ncl*, related with the Stomatology coverage (Appointments and Treatments) per unit of exposure, *exp*.

We are going to assume that claims are generated according to a Homogeneous Poisson Process and then the expected value of the number of claims is proportional to the exposure. The Homogeneous Poisson Process is one of the most implemented processes to model claim counts and is widely exploited in insurance pricing. As we are using the log-link function, we added to the linear predictor the log of the exposure as an offset. The model we implemented in R to estimate the frequency has the following structure:

$glm(ncl \sim ., \text{fam} = poisson(link = \log), \text{offset} = \log(exp), \text{data} = Train)$

As described in the subsection 3.4.1, Variable Selection, we first analysed the significance of each variable by constructing single factor models. With all the potential explanatory factors, that arrised from those models, we constructed a "maximal" model, which was then trimmed down by the application of the R function ***stepAIC***. The model resultant from this procedure was the one that attained the lowest AIC and it comprises 9 factor variables: *age*, *sex*, *fp*, *policy_age*, *fleet*, *fcob*, *dist_company*, *cp2_company* and *bin_pills*.

We have two variables that represent the region where the policyholders work. For

matters of simplicity, we decided to use in R the variable *dist_company* because the *cp2-company* has too many levels, which would be very time consuming to group in R.

Figure 4.1 illustrates, for each variable, the relative frequency of each level in the total number of claims and exposure. By analysing the histograms, we can see that the proportion of the number of claims and of the exposure are similarly distributed through the different levels of each variable, except for the *sex* variable. The majority of the reported claims, (67%), are related to policies that also had coverage for the prescription drugs; the female group corresponds to a greater proportion of the reported claims (54%), however the males correspond to a greater part of the total years of exposure (51%); the monthly payment option is the one selected by most contracts, which represent 57% of the total exposure; we can also see a preference for bank transfer payments; the new policies together with the one-year policies represent 61% of the total years of exposure and 32% of the claims incurred belong to policies with one year of existence; The coverage we are analysing only exists for group policies, hence it is natural to observe the holder as the level with highest proportion of exposure and number of claims. Practically all the policies have Fleet E, they represent 98% of the total exposure.

The *age* and *dist_company* variables are the only multi-level factors with a large number of levels and, in order for them to be included in the model, their levels had to be revised.

To combine the *age* levels, we started by estimating a Poisson GLM for the *ncl* with the *age* as the only covariate and, by an iterative process, we aggregated neighboured levels based on their coefficients and empirical claim frequencies [1]. We obtained the following 7 levels: *[00,04]*, *[05,14]*, *[15,25]*, *[26,38]*, *[39,56]*, *[57,69]* and *[70,99]*, each with a quite satisfactory volume of data and with statistical significance for the model.

From Figure 4.1, we can see that a big part of the total exposure belongs to the most populated cities, Lisboa and Porto, that also correspond to the cities with the major part of the reported claims, as it was expected. Similarly to the *age* variable, *dist_company* had to be grouped. However, this variable does not have an inherent order, so it is more challenging to rearrange the levels. We started by combining the islands into one category and then we analysed the empirical claim frequencies of each district, by an iterative process. We tried to group the closest districts that had similar empirical claim frequencies. However, we also observed some distant districts with similar behaviours.

[1] The empirical claim frequencies of each original age level where we based our grouping process can be found in Appendix A.

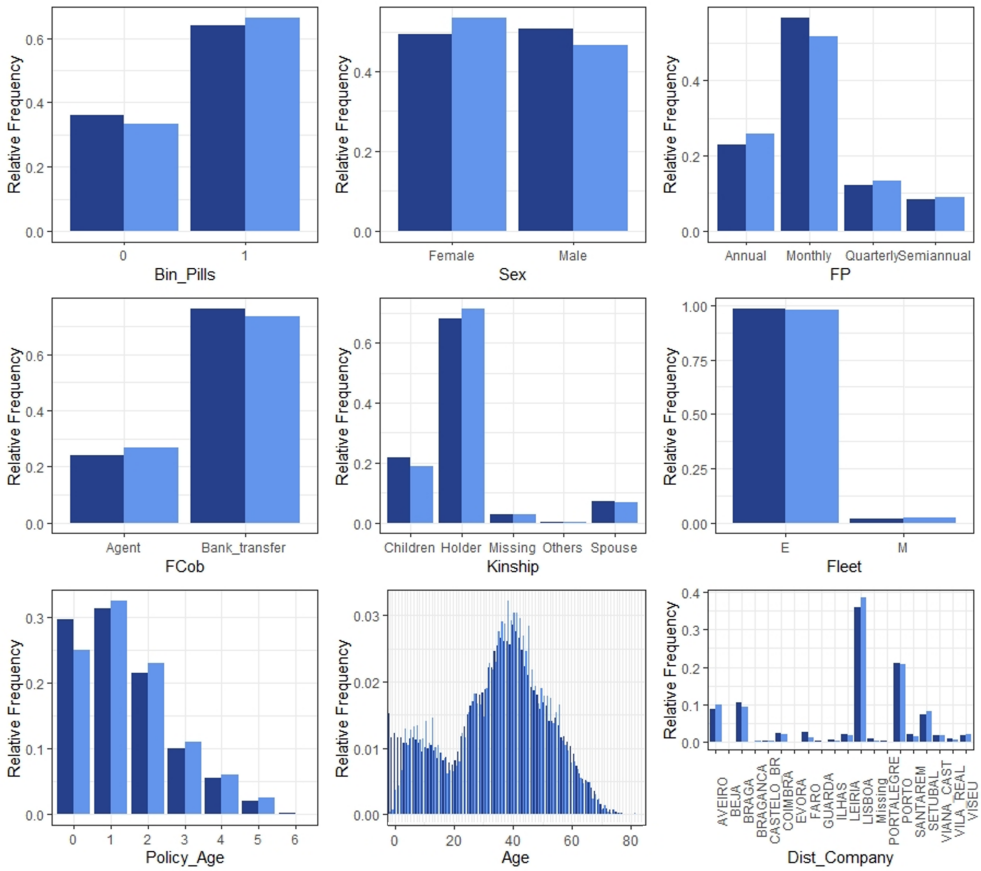

Figure 4.1: Relative frequency of the total years of exposure (dark blue) and the total number of claims (lighter blue) distributed through the different levels of each relevant variable.

We were able to group the 20 initial levels in 9 levels.

By default, R includes in the intercept the first alphabetical level of each factor variable and, since we are dealing with multiplicative models, the coefficients have a relative value towards those levels from the intercept. Therefore, it is relevant to consider as reference level, the level with the highest value of exposure within the same variable. For the *age* factor variable, the model was initially considering the class *[00,04]* in the intercept but we changed it for the class that has the highest proportion of the total exposure: *[39,56]*. The same for the *dist_company*, we changed the order of the factor so the level *Lisboa* could be implemented in the intercept. The variables *bin_pills*, *sex*, *Fp*, *fcob*, *kinship* and *policy_age* also needed to be revised and rearranged.

In Figure A.2, from the Appendix A, we can see the relative frequency of the total years of exposure (dark blue) and the total number of claims (lighter blue) distributed through the different new levels of every variable that needed to be grouped.

After having all the variables properly organized and after reviewing the p-values

of each coefficient, we created several models and compared them by using the cross validation function in R, ***cv.glm***, with k=10.

Once we acquired several good models for the training data, we checked their performance in the hold-out sample. We obtained an optimal final model, with 8 of the explanatory variables previously mentioned, that is displayed in Table 4.1.

We also checked for multicollinearity problems by calculating the variance inflation factors (VIF) for the final model and we did not obtain any problematic value that needed closer attention.

Coefficients	*Estimate*	*Std. error*	*z value*	*Pr(>|z|)*
(Intercept)	-0.847 ***	0.035	-19.408	0.000
age[00,04]	-1.379 ***	0.090	-15.437	0.000
age[05,14]	0.097 *	0.043	2.266	0.023
age[15,25]	-0.133 ***	0.041	-3.231	0.001
age[26,38]	-0.107 ***	0.026	-4.071	0.000
age[57,69]	-0.240 ***	0.041	-5.883	0.000
age[70,99]	-0.433 ***	0.166	-2.614	0.000
sex_Female	0.172 ***	0.021	-8.046	0.000
fp_Annual	0.219 ***	0.027	8.127	0.000
fp_Semiannual	0.217 ***	0.039	5.573	0.000
fp_Quarterly	0.166 ***	0.033	5.027	0.000
policy_age1	0.233 ***	0.028	8.269	0.000
policy_age2	0.266 ***	0.031	8.682	0.000
policy_age3	0.314***	0.038	8.199	0.000
policy_age4	0.266 ***	0.049	5.478	0.000
policy_age5	0.391***	0.071	5.534	0.000
fleet_02_M	-0.361 **	0.074	4.893	0.000
kinship_Spouse_child	-0.109 ***	0.033	-3.301	0.000
dist_Aveiro_Setubal_Vise	0.126 ***	0.030	4.202	0.000
dist_Beja_Evora_Faro	-0.913 ***	0.098	-9.298	0.000
dist_Braga_VilaR_Coimb	-0.126 ***	0.036	-3.507	0.000
dist_Braganca	0.554 **	0.179	3.103	0.002
dist_CastBranc_Portaleg_Guarda	-0.593***	0.149	-3.988	0.000
dist_Ilhas	-0.390 *	0.168	-2.320	0.020
dist_Leiria_Santarem	-0.161 **	0.058	-2.760	0.005
bin_Pills_0	-0.106 ***	0.022	-4.663	0.000

Null Deviance: 52050 *Residual Deviance: 50980* *AIC: 68919* *Significance codes: '***' p < 0.001; '**' p < 0.01; '*' p < 0.05; ' ' p < 0.1;*

Table 4.1: Regression estimates of the Poisson GLM for the Claim Frequency in the training set.

4.2 Claim Severity

For modelling Claim Severity, the target variable we considered was the claim amount per claim incurred, *severity*. To the extent that our dependent variable represents an average amount we comprised the number of claims in the model as weights.

As we already conclude from Table 3.2, 77% of our observations correspond to insured people who did not incurred in a claim during their contract period. Therefore, in order to model the severity, we had to discard those observations and only consider the ones that incurred in losses. The severity distribution is shown in the Figure 4.2. As we can see, the severity has a right highly skewed tail and we can also notice a small peak at the end of the distribution's tail. The reason for this is that 250€ is the upper annual limit covered by the insurance company for claims related with the Stomatology coverage. The

distribution we used for modelling Claim Severity was the Gamma distribution, although we observe an upper limit in the distribution we do not consider the Truncated Gamma for simplicity and because of its minor effect. The model we implemented in R to estimate the severity has the following structure:

$$glm(severity \sim ., \text{fam} = gamma(link = \log), \text{weight} = ncl, \text{data} = Train.sev)$$

Similarly to the Frequency model, we started by implementing the ***stepAIC*** function to observe which variables should be consider in our model building process. The explanatory variables we obtained from this procedure were: *age*, *fp*, *fleet*, *kinship*, *dist_company*, *cp2_company* and *bin_pills*. Because of the complexity of the variable *cp2_company*, we decided once again to use in R the variable *dist_company*.

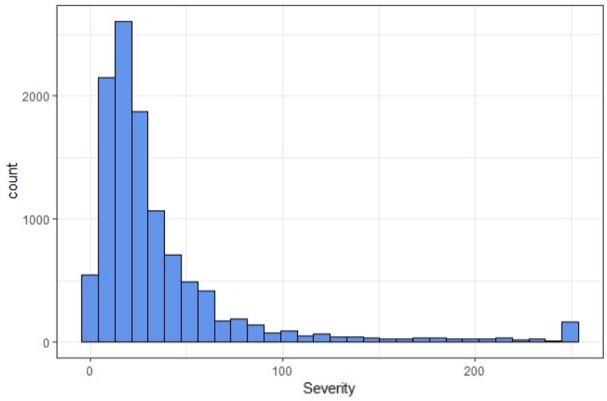

Figure 4.2: Claim cost histogram for the data without the zero claims.

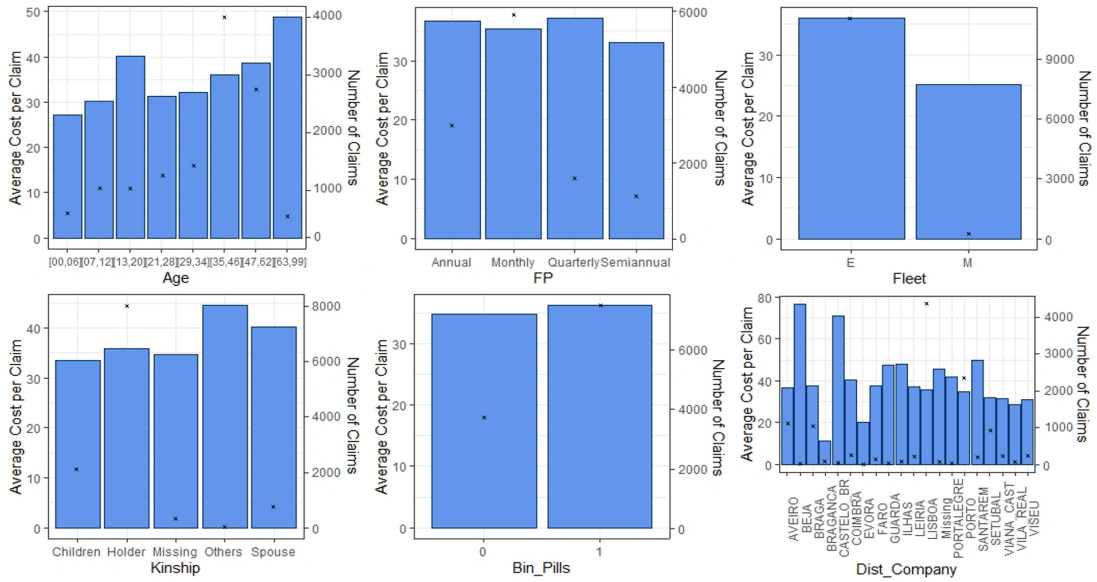

Figure 4.3: The average claim cost (blue bars) and the total number of claims (stars) corresponding to each level of the relevant variables.

The figure 4.3 shows how the average claim cost and the total number of claims vary per level of each variable. Once again, the variables *age* and *dist_company* needed to be grouped in order to decrease their number of levels. The *age* variable is already grouped in the figure 4.3, from where we can see that the older ages, *[63,99]*, have a big average cost when compared to the other age levels. However, they represent a very small number of reported claims, which may explain the high value for the average cost. The age level with the highest number of claims is *[35,46]* that is going to be the level considered in the intercept. For the variables *fp* and *bin_pills*, the average cost are similar between the levels of each variable. It is clear that *Beja* and *Castelo Branco* are the districts that stand out for having a very high average cost but with few number of claims.

When grouping the variables, we proceed in a similar way as described in the Claim Frequency modelling. The empirical claim frequencies we used to help us grouping the *age* variable are represented in Figure A.3, from the Appendix A. For the *dist_company* variable, since some districts had just a few observations, we started by aggregating them to districts with similar behaviour, and only then we started to compare the coefficients for the rest of the districts. The final levels we obtained for each variable used in the final model are shown in Table 4.2.

To contemplate if indeed all the explanatory variables, resultant from the minimization of the AIC, are needed to model the Claim Severity, we created several models and tested them by 10-fold cross validation. We selected the models with the best results in the training set and we ascertained their performance on the test data.

Coefficients	*Estimate*	*Std. error*	*z value*	*Pr(>|z|)*
(Intercept)	3.583 ***	0.026	138.846	0.000
age[00,06]	-0.341 ***	0.067	-5.102	0.000
age[07,12]	-0.170 ***	0.050	-3.372	0.000
age[13,20]	0.125 *	0.050	2.498	0.013
age[21,28]	-0.147 **	0.047	-3.153	0.001
age[29,34]	-0.120 **	0.043	-2.777	0.006
age[47,62]	0.080 *	0.033	2.359	0.019
age[63,99]	0.356 ***	0.073	4.819	0.000
fleet_02_M	-0.377 ***	0.088	-4.262	0.000
fp_Annual	0.056 ·	0.031	1.826	0.068
fp_Semiannual	-0.150 **	0.046	-3.265	0.001
fp_Quarterly	0.080 ·	0.039	1.963	0.050
dist_Beja_CastBranc	0.789 ***	0.221	3.568	0.000
dist_Braganca_Evora	-1.097 ***	0.205	-5.356	0.000
dist_Ilhas_Santarem	0.362 ***	0.087	4.155	0.000
dist_Setubal_Viseu	-0.132 **	0.043	-3.098	0.005
dist_VianaCastelo_VilaR	-0.189 *	0.080	-2.354	0.026

Null Deviance: 8659 *Residual Deviance: 8370* *AIC: 82065* *Significance codes: '***' p < 0.001; '**' p < 0.01; '*' p < 0.05; '' p < 0.1;*

Table 4.2: Regression estimates of the Gamma GLM for the Claim Severity in the training set.

Once we got a satisfying parsimonious model, we analysed the p-values of the coefficients to see if any more levels needed to be grouped with the intercept. We also checked for multicollinearity problems by calculating the VIF. The output we obtained from our

final model in the training set is displayed in Table 4.2.

4.3 Pure Premium

The classical model approach in use allows for the Pure Premium to be calculated by combining the mean estimates from the claim frequency and severity models, generating a multiplicative tariff. Hence, the pure premium corresponds to the expected aggregate claim amount per unit of exposure.

For both models we used the logarithmic link function so we had to take the exponential of each estimated coefficient. In the following equation is presented how to obtain the pure premium from the estimated coefficients for the claim frequency (β_i) and for the claim severity models (α_i) :

$$PurePremium = e^{\beta_0+\alpha_0} \times e^{\beta_1+\alpha_1} \times ... \times e^{\beta_p+\alpha_p}, \tag{4.1}$$

where β_0 and α_0 are the intercept coefficients of the models and the remaining β_i and α_i are the estimated coefficient for the risk factor i, $i = 1, 2, ..., p$.

The premium increase or decrease, associated to each level of risk factor, obtained by combining the two models, is displayed in the Tariff from Table 4.3 and its coefficients are represented in Figure 4.4. To calculate the Stomatology risk premium, for a given policyholder, we have to multiply the base level premium with the Tariff coefficients corresponding to the levels of the variables to which that individual belongs. In Figure 4.4, all levels above the dashed line will increase the premium, meaning that individuals belonging to those levels will pay a higher premium than the ones belonging to levels below the dashed line.

We observe a big decline in the tariff coefficients for the ages *[57,62]* and *[70,99]*. As a marketing strategy for the company, it would be wise to smooth the age effect through the older ages, so that there are no big differences in the premium for someone aged 57 or 62 and someone aged 56 or 63.

The results obtained from modelling the pure premium in R for the Stomatology Appointments and Treatments have some distinctions from the ones obtained by my colleagues. These differences are going to be addressed in the following chapter.

The estimation of the pure premium for the other coverages and sub-coverages was

performed on Emblem. Once the expected losses for all the coverages were estimated, we started to build an automatized simulator. This simulator asks the user what coverages the policyholder wants along with other important information. Afterwards, it automatically calculates the pure premium for a given policy.

Rating Factor Levels		*Rating Factor Levels*		*Rating Factor Levels*	
Base Level	15.422	Aveiro	1.134	Fleet E	1.000
Age [00,04]	0.179	Beja	0.883	Fleet M	0.984
Age [05,06]	0.783	Braga	0.881	Policy age 0	1.000
Age [07,12]	0.930	Bragança	0.581	Policy age 1	1.262
Age [13,14]	1.125	Castelo Branco	1.216	Policy age 2	1.305
Age [15,20]	0.992	Coimbra	0.882	Policy age 3	1.368
Age [21,25]	0.756	Evora	0.600	Policy age 4	1.304
Age [26,28]	0.776	Faro	0.401	Policy age 5	1.478
Age [29,34]	0.797	Guarda	0.553	Policy age 6	1.000
Age [35,38]	0.899	Ilhas	0.973	FP Monthly	1.000
Age [39,46]	1.000	Leiria	0.851	FP Annual	1.316
Age [47,56]	1.083	Lisboa	1.000	FP Semiannual	1.074
Age [57,62]	0.852	Portalegre	0.553	FP Quarterly	1.275
Age [63,69]	1.123	Porto	0.979	Kinship Holder	1.000
Age [70,99]	0.926	Santarem	1.223	Kinship Spouse	0.900
Bin_Pills 1	1.000	Setubal	0.996	Kinship Children	0.900
Bin_Pills 0	0.899	Viana do Castelo	0.810	Kinship Others	1.000
Male	1.000	Vila Real	0.723		
Female	1.188	Viseu	1.996		

Table 4.3: Tariff for the Stomatology Appointments and Treatments model.

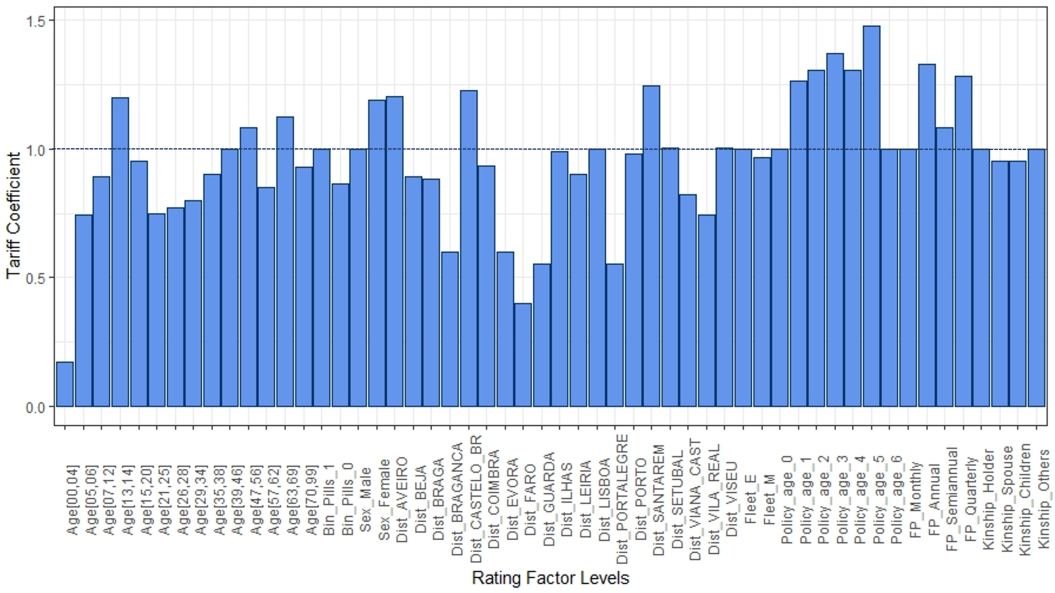

Figure 4.4: Graphical representation of the tariff coefficients from table 4.3.

Chapter 5

Analysis of the Results and Further Developments

5.1 Analysis of the Results

My internship strayed from the initial goals and, because of the home confinement, some arrangements were made. Unfortunately, I could not use Emblem for building the models for the project, but I was always following the process.

We decided that the best approach for my internship report would be for me to model a coverage in R and compare the results with the models that my colleagues obtained with Emblem. For the models constructed by the company, the output is also in a tariff structure, similar to the one displayed in Table 4.3.

We observed some structural differences between the models. Since we do not consider the same set of levels in the intercept, the coefficients obtained by my colleagues suffered some alterations in order to be relative to the same base level and to be compared with the coefficients from R. Another discrepancy resulted from the fact that they did not consider the variable *bin_pills* and treated the *age* variable differently. The model constructed in R used *age* as a categorical variable and grouped it, while the model produced in Emblem considered that variable as a continuous variable and modelled it using a polynomial. To overcome this distinction, we started by considering as the base level the age in the middle of the interval *[39,46]*, which is the level I considered in the intercept. Afterwards, in the Emblem output, we divided each estimated value by the coefficient corresponding to the age 42. Thereupon, with the obtained values we calculated for each age interval the averages of those coefficients. The results obtained can be observed in the Table 5.1.

Another variable that required extra attention was the one considered for the company's region. I used the company's districts in my model, while my colleagues considered the first two digits of the postal code. In order to compare the coefficients from these two variables, we transformed the 2-digit postal code in the respective district. In Figure A.5, from the Appendix A, is displayed the two digits we considered for each Portugal district. Afterwards, we calculated the average of the coefficients per district.

The coefficients obtained for the Pure Premium modeled in Emblem, after performing the described transformations for all the variables, are also displayed in Table 5.1.

Rating Factor	*R Model*	*Emblem Model*	
age[00,04]	0.179	0.205	-13%
age[05,06]	0.783	0.567	-38%
age[07,12]	0.930	0.832	-12%
age[13,14]	1.125	1.120	-12%
age[15,20]	0.992	1.077	8%
age[21,25]	0.756	0.716	-6%
age[26,28]	0.776	0.734	-6%

Rating Factor	*R Model*	*Emblem Model*	
age[29,34]	0.797	0.820	3%
age [35,38]	0.899	0.943	5%
age [39,46]	1.000	1.049	5%
age [47,56]	1.083	1.054	-3%
age [57,62]	0.852	1.057	19%
age [63,69]	1.123	0.965	-16%
age [70,99]	0.926	0.863	-7%

Rating Factor	*R Model*	*Emblem Model*	
Aveiro	1.134	1.011	-12%
Beja	0.883	0.541	-63%
Braga	0.881	0.843	-5%
Bragança	0.581	0.541	-8%
Castelo Branco	1.216	1.011	-20%
Coimbra	0.882	0.863	-2%
Evora	0.600	0.541	-11%
Faro	0.401	0.541	26%
Guarda	0.553	0.541	-2%
Ilhas	0.973	1.006	3%

Rating Factor	*R Model*	*Emblem Model*	
Leiria	0.851	1.000	15%
Lisboa	1.000	1.000	0%
Portalegre	0.553	0.541	-2%
Porto	0.979	1.011	3%
Santarem	1.223	1.000	-22%
Setubal	0.996	1.000	-0%
Viana do Castelo	0.810	0.841	4%
Vila Real	0.723	0.541	-35%
Viseu	0.996	1.011	2%

Rating Factor	*R Model*	*Emblem Model*	
Policy age 0	1.000	1.000	0%
Policy age 1	1.262	1.306	3%
Policy age 2	1.305	1.306	0%
Policy age 3	1.368	1.306	-5%
Policy age 4	1.304	1.306	0%
Policy age 5	1.478	1.306	-13%
Policy age 6	1.000	1.306	23%
FP Monthly	1.000	1.000	0%
FP Annual	1.316	1.307	-1%
FP Semiannual	1.074	1.032	-4%
FP Quarterly	1.275	1.296	2%

Rating Factor	*R Model*	*Emblem Model*	
Male	1.000	1.000	0%
Female	1.188	1.174	-1%
Fleet E	1.000	1.000	0%
Fleet M	0.984	1.061	7%
Kinship Holder	1.000	1.000	0%
Kinship Spouse	0.900	1.080	17%
Kinship Children	0.900	1.080	17%
Kinship Others	1.000	1.000	0%

Table 5.1: Comparison between the coefficients obtained from the two different models.

By analysing Table 5.1, we can see that the bigger difference between the obtained coefficients belongs to the company's district variable. One possible reason for this is that

when we made the transformation from the postal code to the district, we encountered different districts with the same combination for the two first digits of the postal code. Hence, when we calculated the average of the coefficients for each district, we considered coefficients that belong to other regions.

In the *age* variable, the classes *[15,20]*, *[57,62]* and *[63,69]* have coefficients with opposite behaviour towards the base level, which supports the suggestion mentioned regarding smoothing the effect of the older ages. In general, the coefficients for this variable obtained from the R model are smaller. For the remaining variables, no significant differences were found.

We also decided to calculate the risk premium for a set of policyholders with different features, randomly selected, to compare the results obtained by using both models. The premiums we acquired are displayed in Table 5.2.

	Kinship	Sex	Age	FP	Fleet	Policy Age	District	Bin Pills	R	Emblem
1	Holder	M	39	Monthly	E	0	Lisboa	1	15.2	14.6
2	Holder	M	41	Monthly	E	1	Porto	1	18.79	19.15
3	Holder	F	77	Quarterly	E	6	Viseu	1	18.11	17.29
4	Child	F	11	Semiannual	E	0	Leiria	1	18.83	18.19
5	Spouse	M	63	Annual	M	3	Braga	1	25.25	24.61

Table 5.2: The Risk premium obtained by using the models from R and Emblem.

We can notice that, except for policyholder 2, the pure premium resultant from the R model is higher than the one from Emblem. However, despite the consideration of different variables and techniques, the results obtained by using both models were not significantly different for this set of policyholders.

Moreover, during the process we could conclude that the R software is not as straightforward and user-friendly as Emblem, although it also obtains good results. When it comes to manage big amounts of data, SAS is undoubtedly the fastest program.

5.2 Further Developments

Because of the short time frame of the internship and other encountered barriers already discussed, interesting ideas and mechanisms were left out.

Learning the company's software and applying it to model the other coverages would definitely be the next step. Mainly to model the Hospitalization coverage, as it was the only coverage where large losses where considered.

Furthermore, our initial desire was to create a premium simulator, where the value obtained from it would take into consideration the deductibles, capital limits, co-payments and participation rates. Finishing this task would also be considered for the future work.

The traditional approach mentioned in this internship report, Frequency-Severity modelling, is starting to be replaced by more competitive and efficient methods, such as machine learning and artificial intelligence. Overall, they are beginning to have a significant role in insurance pricing. To further develop this work, it would be interesting to first-hand experience the predictive capacity of the machine learning techniques and how to benefit from their application. Another interesting analysis would be to drop the assumption of independence between the Claim Frequency and Claim Severity, this approach is illustrated in the article Garrido et al. (2016).

Bibliography

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Bahnemann, D. (2015). Distributions for actuaries. *CAS monograph series*, *2*.

Dobson, A., & Barnett, A. (2002). Exponential family and generalized linear models. *An Introduction to Generalized Linear Models. New York: Chapman & Hall/CRC*.

Garrido, J., Genest, C., & Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, *70*, 205–215.

Goldburd, M., Khare, A., & Tevet, D. (2016). Generalized linear models for insurance rating. *Casualty Actuarial Society, CAS Monographs Series*, (5).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An introduction to statistical learning with applications in R*. Springer Texts in Statistics.

Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2008). *Modern actuarial risk theory: Using R* (Vol. 128). Springer Science & Business Media.

McCullagh, P., & Nelder, J. (1998). *Generalized linear models*. Chapman & Hall/CRC.

Müller, M. (2012). *Xplore — learning guide: Generalized linear models*. Springer, Berlin, Heidelberg.

Ohlsson E., J. B. (2010). *Non-life insurance pricing with generalized linear models*. Springer, Berlin, Heidelberg.

Pitkänen, P. (1975). Tariff theory. *ASTIN Bulletin: The Journal of the IAA*, *8*(2), 204–228.

Appendix A

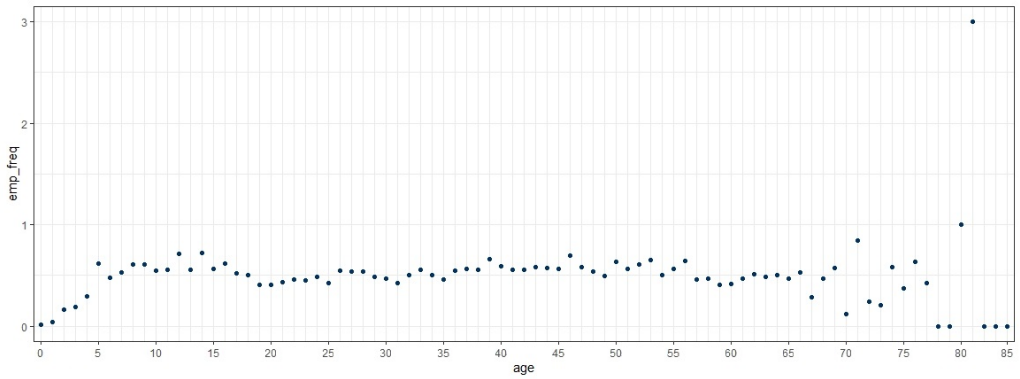

Figure A.1: Empirical claim frequencies of each original age level used for grouping the age levels in the Frequency Model.

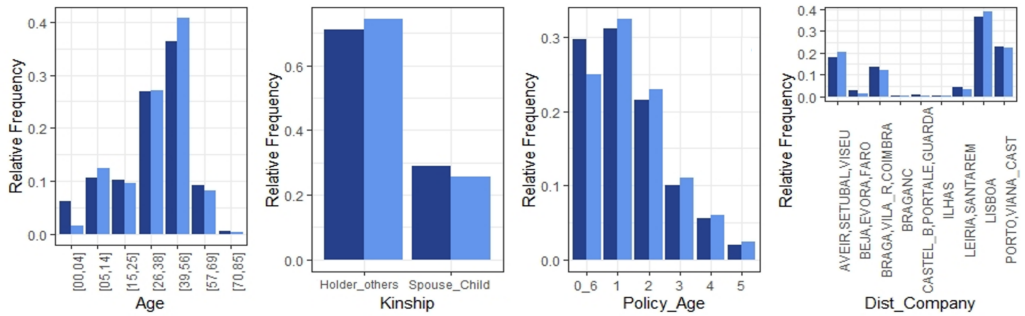

Figure A.2: Relative frequency of the total years of exposure (dark blue) and the total number of claims (lighter blue) distributed through the new levels obtained for the Frequency Model.

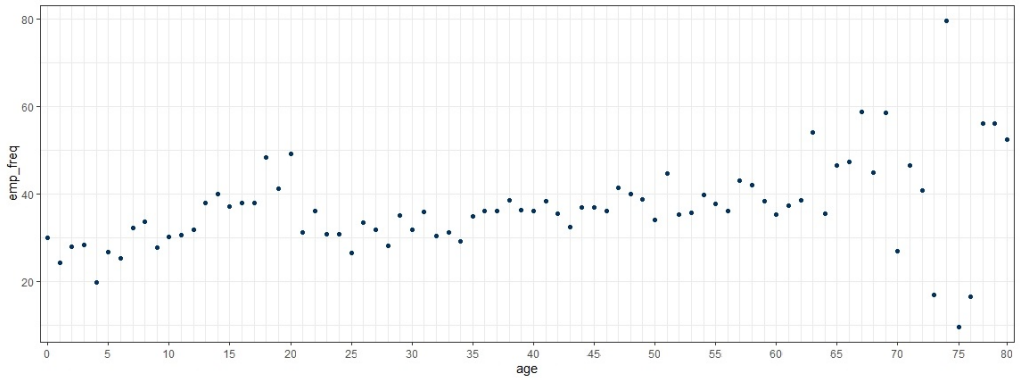

Figure A.3: Empirical claim frequencies of each original age level used for grouping the age levels in the Severity Model.

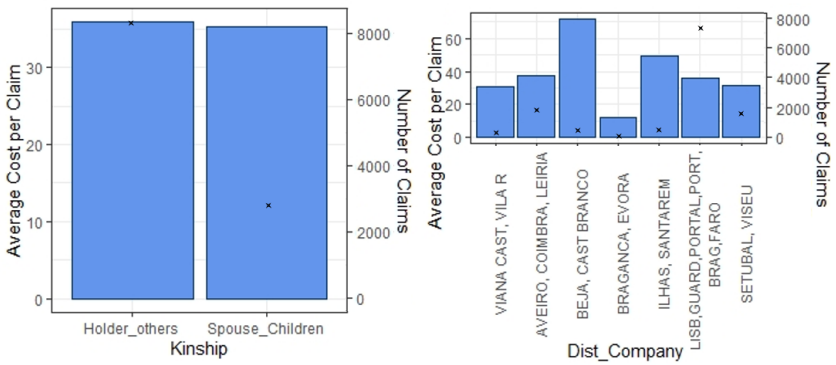

Figure A.4: The average claim cost (blue bars) and the total number of claims (stars) corresponding to the new levels obtained for the Severity Model.

Portugal districts	*The first two postal code digits*
Aveiro	30, 37, 38, 45
Beja	75, 76 ,77 ,78
Braga	46, 47, 48, 49
Bragança	51, 53
Castelo Branco	60, 61, 62
Coimbra	30, 31, 32, 33, 34
Evora	70, 71, 72
Faro	80, 81, 82, 83, 84, 85, 86, 87, 88, 89
Guarda	35, 62, 63, 64
Ilhas	90, 91, 92, 93, 94, 95, 96, 97, 98, 99
Leiria	24, 25, 31, 32
Lisboa	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 26, 27
Portalegre	60, 61, 62, 73, 74, 75
Porto	49, 41, 42, 43, 44, 45, 46, 47, 48
Santarem	20, 21, 22, 23, 24
Setubal	28, 29
Viana do Castelo	49
Vila Real	48, 50, 54
Viseu	34, 35, 36

Figure A.5: The first two digits of the postal code associated with each district.